4К

УДК 004.522

И.С. Кипяткова, А.А. Карпов
Санкт-Петербургский институт информатики и автоматизации РАН, Россия,
kipyatkova@iias.spb.su, karpov@iias.spb.su

Разработка и оценивание модуля транскрибирования для распознавания и синтеза русской речи

Статья описывает модуль транскрибирования, позволяющий создавать базовые транскрипции для слов и текстов, а также альтернативные транскрипции, описывающие вариативность произношения слов в разговорной речи. В статье представлены результаты экспериментов по распознаванию речи с использованием базового и расширенного фонематических словарей.

Введение

Для систем автоматического распознавания и синтеза речи необходимо создавать фонетические транскрипции слов. Транскрипции могут быть созданы вручную, но этот процесс является трудоемким, а при разработке новой системы с другим словарем требуется создавать транскрипции заново. Поэтому предпочтительным является генерировать транскрипции автоматически. Для систем распознавания речи транскрипции создаются по списку слов, а для систем синтеза речи – по входному тексту. При создании транскрипции текста необходимо учитывать фонетические явления, происходящие на стыках слов. А при создании транскрипций для систем распознавания разговорной речи для каждой словоформы необходимо создавать альтернативные транскрипции, которые учитывали бы различные возможные варианты произнесения слов. Поэтому нами был разработан модуль фонематического транскрибирования, который может работать в трех режимах:

1. Создание эталонных транскрипций для списка независимых слов.
2. Создание транскрипций для произвольных связных текстов.
3. Создание альтернативных транскрипций слов, которые учитывают различные варианты возможного произнесения одного и того же слова в разговорной речи.

Первый режим используется для создания базового словаря системы распознавания. В этом режиме на вход модуля транскрибирования поступает список слов, для которых транскрипции создаются с использованием базовых фонетических правил транскрибирования [1] и словаря словоформ с отмеченным ударением (ударениями). При транскрибировании возможны следующие позиционные изменения классов звуков: изменения гласных в положении под ударением, изменения гласных в предударных слогах, изменения гласных в заударных слогах, позиционные изменения согласных. В качестве фонетического алфавита используется модифицированный вариант международного фонетического алфавита SAMPA. В нашем варианте используются 48 фонем: 12 – для гласных звуков (с учетом ударных вариантов) и 36 – для согласных (с учетом твердости и мягкости звуков). Знак [!] используется для обозначения ударения в слове, знак [`] – для обозначения второстепенного ударения и знак ['] – для обозначения мягкости согласных. Алгоритм создания базовых транскрипций слов описан в [2]. Полученные транскрипции затем могут быть использованы для второго и третьего режимов работы модуля. Предва-

рительным этапом создания транскрипций является определение положения ударения в слове. В следующем разделе будет рассмотрен процесс нахождения ударения в слове, а в последующих разделах будут представлены режимы создания транскрипций для текстов и альтернативных транскрипций слов. В последнем разделе будут представлены результаты распознавания речи с использованием различных словарей.

Определение положения ударения в слове

Для создания транскрипций слов необходимо наличие базы данных словоформ русского языка с отметкой ударения. В качестве такой базы использовались две базы данных, доступные в Интернете: (1) – созданная в ходе проекта STARLING (руководитель проекта С.А. Старостин) [3]; (2) – являющаяся частью морфологического анализатора, разработанного А.В. Сокирко [4]. Первая база данных содержит около 1 млн 800 тыс. различных словоформ, это количество словоформ является недостаточным для описания русского языка. В этой базе для некоторых сложных слов проставлено второстепенное ударение. Вторая база данных содержит свыше 2 млн 200 тыс. словоформ. Однако в этой базе данных, в отличие от первой, отсутствует буква *ё* и информация о второстепенном ударении. Поэтому эти две базы данных были объединены, объем получившейся базы данных превысил 2 млн 300 тыс. различных словоформ.

Блок-схема алгоритма простановки ударений для исходной словоформы представлена на рис. 1. В служебных словах (предлоги, союзы), состоящих из одного слога, гласная является безударной. Поскольку для автоматического транскрибирования текста необходима информация о положении ударной гласной, то для служебных слов транскрипции были созданы вручную. Если слово является знаменательным, то положение ударной гласной определяется по получившейся базе данных. Однако в этой базе данных для многих сложных слов отсутствует второстепенное ударение, поэтому если для исходного слова в базе данных отмечено два ударения, то основное и второстепенное ударения проставляются в соответствии с тем, как указано в базе данных. Если же для исходного слова не отмечено два ударения, то осуществляется проверка, является ли слово сложным. Для этого сначала производится проверка, есть ли в слове дефис. Если слово написано через дефис, тогда это слово разбивается на две части, и затем эти две части слова по отдельности ищутся в базе данных ударений. Если они обнаруживаются в базе данных, второстепенное ударение ставится на первое слово, а основное – на второе. Если отдельных частей слова в базе данных нет, но есть исходная словоформа, у которой отмечено одно ударение, то тогда в исходной словоформе ставится основное ударение в соответствии с базой данных. Если исходное слово не содержит дефиса, тогда осуществляется проверка, является ли начало слова префиксом иноязычного происхождения (например, *псевдо-*, *анти-*, *квази-*).

Если начало слова содержится в списке иноязычных префиксов, то происходит поиск оставшейся части слова в базе данных словоформ. Если начало слова не найдено в списке префиксов, или конец слова не найден в базе данных словоформ, то осуществляется поиск этого слова целиком в базе данных словоформ, и ударение ставится в соответствии с тем, как указано в ней. Во второй базе данных словоформ вместо буквы *ё* употребляется *е*, поэтому если целиком слово в ней не найдено, то происходит проверка, есть ли в исходном слове буква *ё*. Если буква *ё* есть, то ударение ставится на эту букву (справедливо всегда). Если буквы *ё* в исходном слове нет и по базе данных ударение также не найдено, то это слово не транскрибируется из-за невозможности корректно проставить ударение.

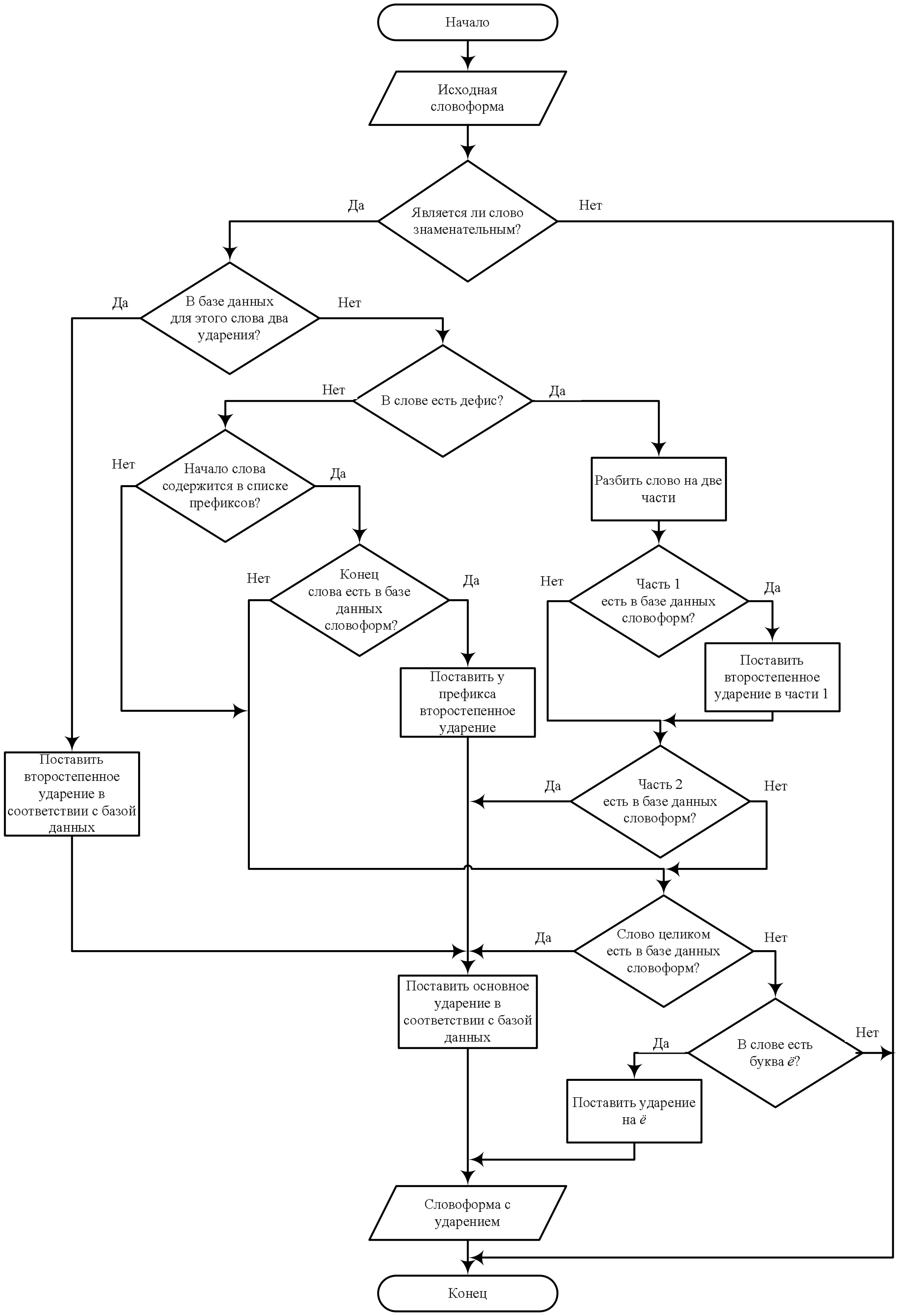


Рисунок 1 – Блок-схема алгоритма простановки ударений для слова

Создание транскрипций для текстов

Данный режим работы может использоваться в синтезаторе речи и при обучении систем распознавания речи. В этом режиме на вход модуля транскрибирования поступают не отдельные слова, а тексты. При транскрибировании текстов учитывается лексический контекст, то есть фонетические явления, происходящие в слитной речи на границах слов, поскольку при слитном произношении начала и концы слов зависят от соседних слов.

При транскрибировании текстов для стыков слов применяется ряд правил, описывающих межсловные фонетические явления [5]:

1. Если в начале слова стоит сочетание фонем /йи/, причем гласная безударная, оно переходит в фонему /ы/ в случае, если первое слово заканчивается на твердую согласную (*город в Якутии* /го!рат в йику!т'ии/ → /го!рат в ыку!т'ии/).
2. Первая в слове гласная /и/ после всех твердых согласных переходит в фонему /ы/ (*лист ивы* /л'и!ст ы!вы/).
3. Безударные гласные редуцируются до полного исчезновения, если они находятся:

   a) между одинаковыми согласными (*мясо сырое* /м'а!са сыро!йе/ → /м'а!с сыро!йе/);

   b) после одной из парных по глухости – звонкости согласных и перед соответствующей парной согласной (*степи большие* /с'т'е!пи бал'шы!йе/ → /с'т'е!п' бал'шы!йе/).
4. Фонемы /т'/ и /д'/, стоящие после /с'/ и /з'/ соответственно, редуцируются до полного исчезновения (*есть порох* /йэ!с'т' по!рах/ → /йэ!с' по!рах/).
5. Фонемы /т/ и /д/, стоящие после /с/ и /з/ соответственно, редуцируются до полного исчезновения (*хвост коровы* /хво!ст каро!вы/ → /хво!с каро!вы/).
6. Согласная /й/ в конце слова редуцируется до полного исчезновения, если ей предшествует безударная гласная, а следующее слово начинается с любой фонемы, кроме ударной гласной (*красный шар* /кра!сный ша!р/ → /кра!сны ша!р/).
7. На стыке двух знаменательных слов глухие согласные /п/, /п'/, /т/, /т'/, /к/, /к'/, /ф/, /ф'/, /с/, /с'/, /ш/, /ш/, /ц/, /ч/ озвончаются перед фонемами /б/, /д/, /г/, /з/ или /ж/. На стыке служебного и знаменательного слова внутрисловные правила ассимиляции по глухости – звонкости сохраняются, т.е. в положении перед глухими шумными согласными звонкие шумные согласные оглушаются, и на их месте выступают глухие шумные, в положении перед звонкими шумными согласными, кроме /в/, /в'/, глухие шумные озвончаются, и на их месте выступают звонкие шумные (*с дороги* /здаро!г'и/, *в лесу* /вл'эсу!/).
8. Сочетание фонем /с'т'/ в конце слова переходит в фонему /щ/, если следующее слово начинается с /ч/ (*есть чему* /йэ!с'т' чэму!/ → /йэ!щ чэму!/).
9. Если на стыке двух слов находятся одинаковые согласные, то согласная первого слова редуцируется (*лес сосновый* /л'э!с сасно!вый/ → /л'э! сасно!вый/).

При обработке текста учитываются знаки препинания. Поскольку на знаках препинания люди обычно делают паузу, стыки слов, разделенных каким-либо знаком препинания, рассматриваются без контекста соседнего слова.

Создание альтернативных транскрипций

Альтернативные транскрипции необходимы при разработке систем распознавания разговорной речи. В разговорной речи произношение слов варьируется: различные дикторы могут произносить одно и то же слово по-разному, кроме того, произношение одного и того же диктора может меняться в зависимости от контекста и темпа речи. Для разговорного стиля речи характерны такие явления, как ассимиляция, а также редукция некоторых фонем вплоть до полного исчезновения. Поэтому транскрипции произнесенных слов часто не совпадают с транскрипциями, сделанными по фонетическим правилам

русского языка. Например, слово *шестьдесят*, которое имеет базовую транскрипцию /ш ы з' д' и с' а! т/, в разговорной речи часто произносится как /ш ы с' а! т/ или даже /ш с' а! т/. Для учета явлений редукции и ассимиляции необходимо добавление альтернативных транскрипций в словарь системы распознавания.

Разработанный модуль фонематического транскрибирования создает альтернативные транскрипции, используя правила, описывающие возможные явления редукции и ассимиляции фонем [5], [6]. Алгоритм автоматического создания альтернативных транскрипций и расширенного словаря системы распознавания представлен на рис. 2.

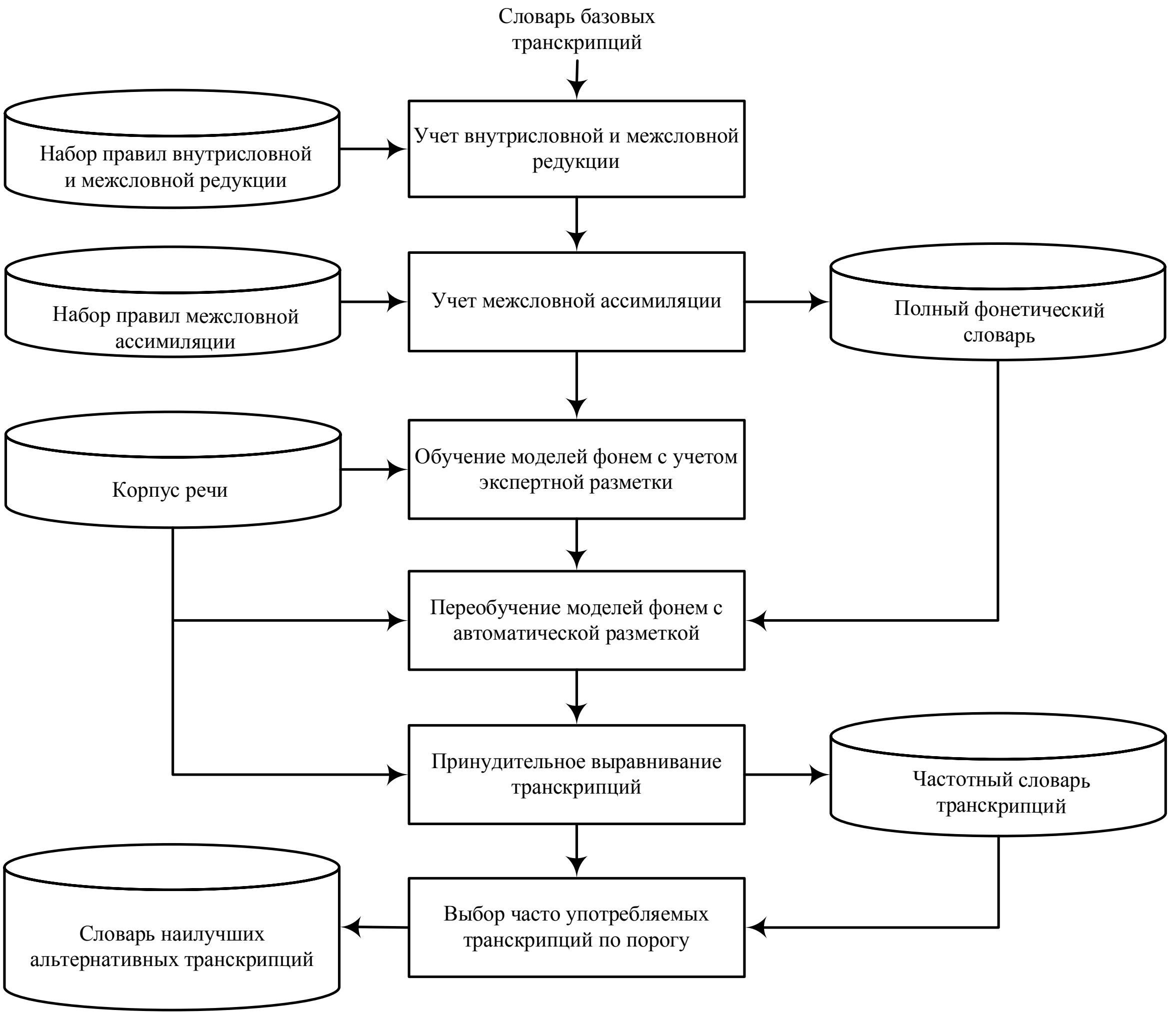


Рисунок 2 – Диаграмма процесса создания расширенного словаря системы распознавания речи

Базовые транскрипции слов поступают в блок учета внутрисловной и межсловной редукции, где для каждой базовой транскрипции слова определяется, какие фонемы подвержены редукции. На выходе блока получается набор альтернативных транскрипций данной словоформы, учитывающий все возможные сочетания редуцируемых фонем. Далее в блоке учета межсловной ассимиляции производится анализ первых и последних фонем в транскрипции, в случае обнаружения фонем, подверженных ассимиляции, производится генерация всех возможных контекстно-зависимых вариантов транскрипций. Таким образом, получается полный словарь всех возможных альтернативных транскрипций.

В качестве акустических моделей речевых единиц используются скрытые Марковские модели (СММ) с тремя состояниями [7]. Начальное создание и обучение моделей фонем производится с учетом экспертной разметки части речевого корпуса, а затем модели дополнительно обучаются с использованием автоматической разметки. Для выбора из множества альтернативных транскрипций осуществляется их принудительное выравнивание (forced alignment), при котором распознаватель выбирает из списка альтернативных транскрипций наиболее подходящую речевому сигналу и сегментирует сигнал на фонемы с их временными метками. В этом случае выбор транскрипции происходит только между альтернативными транскрипциями одного и того же слова, а не между транскрипциями разных слов [8].

Наилучшая транскрипция выбирается следующим образом [9]:

$$\hat{B} = \arg\max_{B} P(B \mid A, W) = \arg\max_{B} P(A, B \mid W) = \arg\max_{B} P_{AM}(A \mid B, W) P_D(B \mid W),$$

где $\hat{B}$ находится по алгоритму Витерби [7]; P_{AM}, P_D представляют собой основные акустические модели и словарь соответственно, A – последовательность векторов признаков, B – последовательность фонем, W – последовательность произнесенных слов. Альтернативные транскрипции рассматриваются равновероятными.

Основой процедуры принудительного выравнивания является алгоритм Витерби, который находит оптимальную последовательность состояний СММ на основе максимальной акустической вероятности. Суть классического алгоритма Витерби заключается в задании начальных параметров модели с последующим чередованием фаз оценки и максимизации данных параметров по критерию максимального правдоподобия.

Для каждого выравнивания алгоритм Витерби вычисляет вероятность того, что фонематическая транскрипция и речевой сигнал подходят друг другу. Наибольшие вероятности при выравнивании транскрипций каждого слова позволяют выбрать оптимальные варианты транскрипций. В результате выполнения принудительного выравнивания выбирается транскрипция, наиболее оптимально подходящая определенному участку речевого сигнала. Транскрипции, которые ни разу не выбрались при принудительном выравнивании, исключаются из словаря, и таким образом создается сокращенный словарь транскрипций. Однако этот сокращенный словарь является избыточным и содержит редкие варианты произношения, что приводит к увеличению акустической и лексической неоднозначности. Поэтому для уменьшения избыточности словаря производится анализ того, насколько часто каждая альтернативная транскрипция выбиралась в ходе обучения, и создается частотный словарь транскрипций. Таким образом, в итоговый расширенный словарь добавляются только те транскрипции, относительная частота появления которых выше определенного порога. В результате создается расширенный (относительно базового) словарь фонематических транскрипций, содержащий наилучшие транскрипции для каждого слова.

Результаты экспериментов

Из табл. 1, в которой представлен фрагмент транскрибированного текста, можно увидеть различия между базовой транскрипцией и транскрипцией, учитывающей межсловные фонетические явления. В слове *вокруг* конечная фонема /к/ была редуцирована, поскольку следующее слово (*которого*) начинается с фонемы /к/. Аналогично происходит редукция фонемы /м/ в слове *снегом*. На стыке слов *дерево* и *вокруг* не происходит редукции фонемы /а/, находящейся между фонемами /в/, поскольку эти слова разделены запятыми. В слове *засыпанный* согласно правилу 6 редуцируется фонема /й/, поскольку

перед ней стоит безударная гласная, а следующее слово начинается не с ударной гласной. В слове же *гигантский* редукция /й/ не происходит, потому что между словами *гигантский* и *засыпанный* стоит запятая.

Таблица 1 – Пример фонематического транскрибирования предложения

Исходный текст	Базовая транскрипция	Транскрипция, учитывающая межсловные явления
Случайно взгляд мой упал на дерево, вокруг которого расположился гигантский, засыпанный снегом муравейник.	случа!йна взгл'а!т мо!й упа!л на д'э!р'ева вакру!к като!рава распалажы!лс'а г'ига!нск'ий засы!паный сн'э!гам мурав'э!йн'ик	случа!йна взгл'а!т мо!й упа!л на д'э!р'ева вакру! като!рава распалажы!лс'а г'ига!нск'ий засы!паны сн'э!га мурав'э!йн'ик

Для оценки разработанного модуля транскрибирования были проведены эксперименты по распознаванию слов и слитно произнесенных фраз при использовании базового и расширенного словаря. Для обучения и тестирования системы распознавания был выбран речевой корпус, содержащий записи произнесенных различными дикторами семизначных номеров телефонов, таким образом длина фразы варьировалась от трех до семи слов. Запись корпуса производилась по аналоговому телефонному каналу с частотой дискретизации 11 кГц, 16 бит на отсчет, моно. Всего корпус содержит около 1000 фраз, из них 80% фраз каждого диктора использовались для обучения системы и 20% – для тестирования. В записи корпуса приняли участие 32 диктора, их средний возраст составил 22 года. Для распознавания слитной русской речи использовался разработанный в СПИИРАН декодер SIRIUS [10], основанный на представлении словаря распознавания в виде двухуровневого морфофонемного префиксного графа. Результаты распознавания представлены в табл. 2.

Таблица 2 – Результаты распознавания речи при различных способах создания фонематических транскрипций для словаря

	Транскрипции, составляющие словарь			
	базовые, созданные автоматически	все альтернативные, созданные автоматически	выбранные при обучении	с порогом 0,15 для альтернативных
Количество транскрипций	37	264	181	75
Ошибка распознавания слов, %	3,92	3,79	3,65	3,38
Ошибка распознавания фраз, %	12,99	12,43	11,86	10,17

При распознавании с базовым словарем объемом в 37 слов количество неправильно распознанных слов составило 3,92%, количество неправильно распознанных фраз – 12,99%. После применения правил редукции и ассимиляции объем словаря увеличился по отношению к базовому более чем в 7 раз и составил 264 транскрипции. Точность распознавания увеличилась по отношению к точности распознавания с базовыми транскрипциями на 0,13% по словам и на 0,56% по фразам. После исключения из словаря тех транскрипций, которые ни разу не выбрались при обучении, объем словаря составил 181 транскрипцию. При этом точность распознавания немного выросла. Затем был введен порог, равный минимально-допустимой относительной частоте встречаемости каждой транскрипции в обучающем корпусе. Наибольшая точность распознавания была достигнута при пороге 0,15: по словам 96,62% и по фразам 89,83%. При данном пороге для каждого слова в среднем было по 2,03 транскрипции.

Выводы

Разработанный модуль транскрибирования позволяет создавать фонематические транскрипции как для списка слов, так и для текстов. Создание транскрипций для текстов является особенно важным для систем автоматического синтеза речи, поскольку полученные транскрипции описывают фонетические явления, происходящие на стыках слов. Также данный модуль транскрибирования позволяет создавать альтернативные транскрипции слов, учитывающие явления редукции и ассимиляции, возникающие в разговорной речи. Расширенный словарь с альтернативными транскрипциями может быть использован для систем распознавания разговорной речи. Проведенные эксперименты по распознаванию речи с использованием базового и расширенного словаря показывают, что использование альтернативных транскрипций увеличивает точность распознавания. Однако использование слишком большого числа альтернативных транскрипций увеличивает лексическую неоднозначность и может привести к снижению точности распознавания. Поэтому необходимо ограничивать число альтернативных транскрипций путем введения порога. Регулируя величину порога, можно повысить точность распознавания. Для словаря, использованного в данной работе, оптимальный порог был равен 0,15. В дальнейшем планируется проверка работы данного модуля транскрибирования для словаря большого объема.

Работа проводится при поддержке фонда РФФИ: проект № 08-08-00128-а «Моделирование нефонемных речевых элементов и создание альтернативных транскрипций для распознавания спонтанной русской речи» и проект № 09-07-91220-СТ_а «Методы и многомодальные интерфейсы для бесконтактной коммуникации инвалидов с информационно-справочными системами».

Литература

1. Русская грамматика : в 2 т. / Редкол. : Н.Ю. Шведова (гл. ред.)[и др.]. – М. : Наука, 1980. – 783 с.
2. Кипяткова И.С. Модуль фонематического транскрибирования для системы распознавания разговорной русской речи / И.С. Кипяткова, А.А. Карпов // Искусственный интеллект. – 2008. – № 4. – С. 747-757.
3. Режим доступа : http://starling.rinet.ru/morpho.php?lan=ru
4. Режим доступа : http://www.aot.ru/
5. Лобанов Б.М. Моделирование внутрисловных и межсловных фонетико-акустических явлений полного и разговорного стилей в системе синтеза речи по тексту / Б.М. Лобанов, Л.И. Цирульник // Труды Первого междисциплинарного семинара «Анализ разговорной русской речи» ($АР^3$-2007). – СПб. : ГУАП, 2007. – С. 57-71.
6. Русская разговорная речь / под редакцией Е.А. Земской. – М. : Наука, 1973. – 485 с.
7. Rabiner L. Fundamentals of Speech Recognition. Prentice Hall / L. Rabiner, B.-H. Juang. – 1995. – 507 p.
8. Kessens J. M. Improving the performance of Dutch CSR by modeling within-word and cross-word pronunciation variation / J.M. Kessens, M. Wester, H. Strik // Speech Communication. – 1999. – Vol. 29. – P. 193-207.
9. Saraclar M. Pronunciation Modeling for Conversational Speech Recognition / M. Saraclar // PhD thesis. – Baltimore, USA, 2000.
10. Ronzhin A. Morpho-Phonetic Tree Decoder for Russian / [Ronzhin A., Leontieva An., Kagirov I., Carpov A.] // Proc. of 12-th International Conference on Speech and Computer SPECOM. – Moscow (Russia), 2007. – P. 491-498.

І.С. Кип'яткова, О.А. Карпов

Розробка і оцінювання модуля транскрибування для розпізнавання і синтезу російської мови

Стаття описує модуль транскрибування, що дозволяє створювати базові транскрипції для слів та текстів, а також альтернативні транскрипції, які описують варіативність вимови у розмовній мові. У статті наявні результати експериментів з розпізнавання мови з використанням базового та розширеного фонематичних словників.

I.S. Kipyatkova, A.A. Karpov

Development and Evaluation of the Transcription Module for Recognition and Synthesis of Russian Speech

The paper describes the transcription module that allows creating basic transcriptions for words and texts as well as alternative transcriptions that describe word pronunciation variability in conversational speech. Experimental results on automatic speech recognition with basic and extended phonemic dictionaries are presented in the paper.

Статья поступила в редакцию 14.07.2009.